Optionen und Architekturen für Augmented Generation abrufen auf AWS

AWS Prescriptive Guidance

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

AWS Prescriptive Guidance: Optionen und Architekturen für Augmented Generation abrufen auf AWS

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Die Handelsmarken und die Handelsaufmachung von Amazon dürfen nicht in einer Weise in Verbindung mit nicht von Amazon stammenden Produkten oder Services verwendet werden, durch die Kunden irregeführt werden könnten oder Amazon in schlechtem Licht dargestellt oder diskreditiert werden könnte. Alle anderen Handelsmarken, die nicht Eigentum von Amazon sind, gehören den jeweiligen Besitzern, die möglicherweise zu Amazon gehören oder nicht, mit Amazon verbunden sind oder von Amazon gesponsert werden.

Table of Contents

Einführung 1
Zielgruppe 1
Ziele 2
Generative KI-Optionen 3
RAG verstehen 5
Komponenten 6
Vergleich von RAG und Feinabstimmung 7
Anwendungsfälle für RAG 10
Vollständig verwaltete RAG-Optionen 11
Wissensdatenbanken für Amazon Bedrock 11
Datenquellen 13
Vektor-Datenbanken 15
Amazon Q Business 16
Schlüssel-Features 16
End-user Anpassung 18
Amazon SageMaker AI-Leinwand 18
Kundenspezifische RAG-Architekturen 21
Retriever 21
Amazon Kendra 22
OpenSearch Amazon-Dienst 24
Amazon Aurora PostgreSQL und pgvector 24
Amazon Neptune Analytics 25
Amazon MemoryDB 26
Amazon DocumentDB 27
Tannenzapfen 29
MongoDB-Atlas 30
Weben 31
Generatoren 32
Amazon Bedrock 32
SageMaker AI JumpStart 33
Auswahl einer RAG-Option 34
Schlussfolgerung 36
Dokumentverlauf 37
..... xxxviii

Optionen und Architekturen für Augmented Generation abrufen auf AWS

Mithil Shah, Rajeev Muralidhar und Natacha Fort, Amazon Web Services

Oktober 2024 (Geschichte der Dokumente)

Generative KI bezieht sich auf eine Untergruppe von KI-Modellen, die mit einer einfachen Texteingabe neue Inhalte und Artefakte wie Bilder, Videos, Text und Audio erstellen können. Generative KI-Modelle werden anhand riesiger Datenmengen trainiert, die eine Vielzahl von Themen und Aufgaben umfassen. Dadurch können sie eine bemerkenswerte Vielseitigkeit bei der Ausführung verschiedener Aufgaben unter Beweis stellen, auch bei solchen, für die sie nicht explizit geschult wurden. Aufgrund der Fähigkeit eines einzelnen Modells, mehrere Aufgaben auszuführen, werden diese Modelle oft als Basismodelle (FMs) bezeichnet.

Eine der bemerkenswerten Anwendungen generativer KI-Modelle ist ihre Fähigkeit, Fragen zu beantworten. Es gibt jedoch spezifische Herausforderungen, die sich ergeben, wenn diese Modelle zur Beantwortung von Fragen auf der Grundlage benutzerdefinierter Dokumente verwendet werden. Benutzerdefinierte Dokumente können firmeneigene Informationen, interne Websites, interne Dokumentation, Confluence Seiten, SharePoint Seiten und andere enthalten. Eine Option ist die Verwendung von Retrieval Augmented Generation (RAG). Bei RAG verweist das Foundation-Modell vor der Generierung einer Antwort auf eine autoritative Datenquelle, die sich außerhalb der Trainingsdatenquellen befindet (z. B. Ihre benutzerdefinierten Dokumente).

In diesem Leitfaden werden die verschiedenen generativen KI-Optionen beschrieben, die für die Beantwortung von Fragen aus benutzerdefinierter Dokumentation zur Verfügung stehen, einschließlich Retrieval Augmented Generation (RAG) -Systemen. Es bietet auch einen Überblick über die Erstellung von RAG-Systemen auf Amazon Web Services (AWS). Wenn Sie sich die RAG-Optionen und -Architekturen ansehen, können Sie zwischen vollständig verwalteten Services auf AWS und benutzerdefinierten RAG-Architekturen wählen.

Zielgruppe

Die Zielgruppe dieses Leitfadens sind Architekten und Manager generativer KI, die eine RAG-Lösung entwickeln, die verfügbaren Architekturen überprüfen und die Vor- und Nachteile der einzelnen Optionen verstehen möchten.

Ziele

Dieser Leitfaden hilft Ihnen bei folgenden Aufgaben:

- Machen Sie sich mit den generativen KI-Optionen vertraut, die zur Beantwortung von Fragen aus benutzerdefinierten Dokumenten zur Verfügung stehen
- Sehen Sie sich die Architekturoptionen für RAG-Systeme an unter AWS
- Machen Sie sich mit den Vor- und Nachteilen der einzelnen RAG-Optionen vertraut
- Wählen Sie eine RAG-Architektur für Ihre AWS Umgebung

Generative KI-Optionen für die Abfrage benutzerdefinierter Dokumente

Organizations verfügen häufig über verschiedene Quellen für strukturierte und unstrukturierte Daten. Dieser Leitfaden konzentriert sich darauf, wie Sie generative KI verwenden können, um Fragen aus unstrukturierten Daten zu beantworten.

Unstrukturierte Daten in Ihrem Unternehmen können aus verschiedenen Quellen stammen. Dies können PDFs, Textdateien, interne Wikis, technische Dokumente, öffentlich zugängliche Websites, Wissensdatenbanken oder andere sein. Wenn Sie ein Basismodell benötigen, das Fragen zu unstrukturierten Daten beantworten kann, sind die folgenden Optionen verfügbar:

- Trainieren Sie ein neues Basismodell, indem Sie Ihre benutzerdefinierten Dokumente und andere Trainingsdaten verwenden
- Fine-tune ein vorhandenes Basismodell mithilfe von Daten aus Ihren benutzerdefinierten Dokumenten
- Verwenden Sie kontextbezogenes Lernen, um ein Dokument an das Foundation-Modell weiterzugeben, wenn Sie eine Frage stellen
- Verwenden Sie einen RAG-Ansatz (Retrieval Augmented Generation)

Es ist ein ehrgeiziges Unterfangen, ein neues Basismodell von Grund auf neu zu entwickeln, das Ihre benutzerdefinierten Daten enthält. Einige Unternehmen haben dies erfolgreich getan, beispielsweise Bloomberg mit ihrem BloombergGPTModell. Ein anderes Beispiel ist das multimodale EXAONEModell vonLG AI Research, das anhand von 600 Milliarden Kunstwerken und 250 Millionen hochauflösenden Bildern mit Text trainiert wurde. Laut The Cost of AI: Should You Build or Buy Your Foundation Model (LinkedIn) Meta Llama 2 kostet die Schulung eines ähnlichen Modells rund 4,8 Millionen US-Dollar. Es gibt zwei Hauptvoraussetzungen, um ein Modell von Grund auf zu trainieren: Zugang zu Ressourcen (finanzielle, technische, zeitliche) und eine klare Investitionsrendite. Wenn dies nicht die richtige Lösung zu sein scheint, besteht die nächste Option darin, ein vorhandenes Basismodell zu verfeinern.

Fine-tuning Bei einem vorhandenen Modell wird ein Modell, z. B. ein Amazon Titan-, Mistral- oder Lama-Modell, verwendet und das Modell anschließend an Ihre benutzerdefinierten Daten angepasst. Es gibt verschiedene Techniken zur Feinabstimmung, von denen die meisten nur die Änderung einiger weniger Parameter beinhalten, anstatt alle Parameter im Modell zu ändern. Dies

wird als parametereffiziente Feinabstimmung bezeichnet. Es gibt zwei Hauptmethoden für die Feinabstimmung:

- Die überwachte Feinabstimmung verwendet beschriftete Daten und hilft Ihnen, das Modell für eine neue Art von Aufgabe zu trainieren. Wenn Sie beispielsweise einen Bericht auf der Grundlage eines PDF-Formulars erstellen möchten, müssen Sie dem Modell möglicherweise anhand von ausreichend Beispielen beibringen, wie das geht.
- Die unbeaufsichtigte Feinabstimmung ist aufgabenunabhängig und passt das Basismodell an Ihre eigenen Daten an. Es trainiert das Modell, den Kontext Ihrer Dokumente zu verstehen. Das fein abgestimmte Modell erstellt dann Inhalte, z. B. einen Bericht, und verwendet dabei einen Stil, der besser an Ihre Organisation angepasst ist.

Eine Feinabstimmung ist jedoch möglicherweise nicht ideal für Anwendungsfälle mit Fragen und Antworten. Weitere Informationen finden Sie unter Vergleich von RAG und Feinabstimmung in diesem Handbuch.

Wenn Sie eine Frage stellen, können Sie einem Dokument das Grundlagenmodell übergeben und das kontextbezogene Lernen des Modells nutzen, um Antworten aus dem Dokument zurückzugeben. Diese Option eignet sich für die Ad-hoc-Abfrage eines einzelnen Dokuments. Diese Lösung eignet sich jedoch nicht gut für die Abfrage mehrerer Dokumente oder für die Abfrage von Systemen und Anwendungen wie Microsoft SharePoint oder Atlassian Confluence.

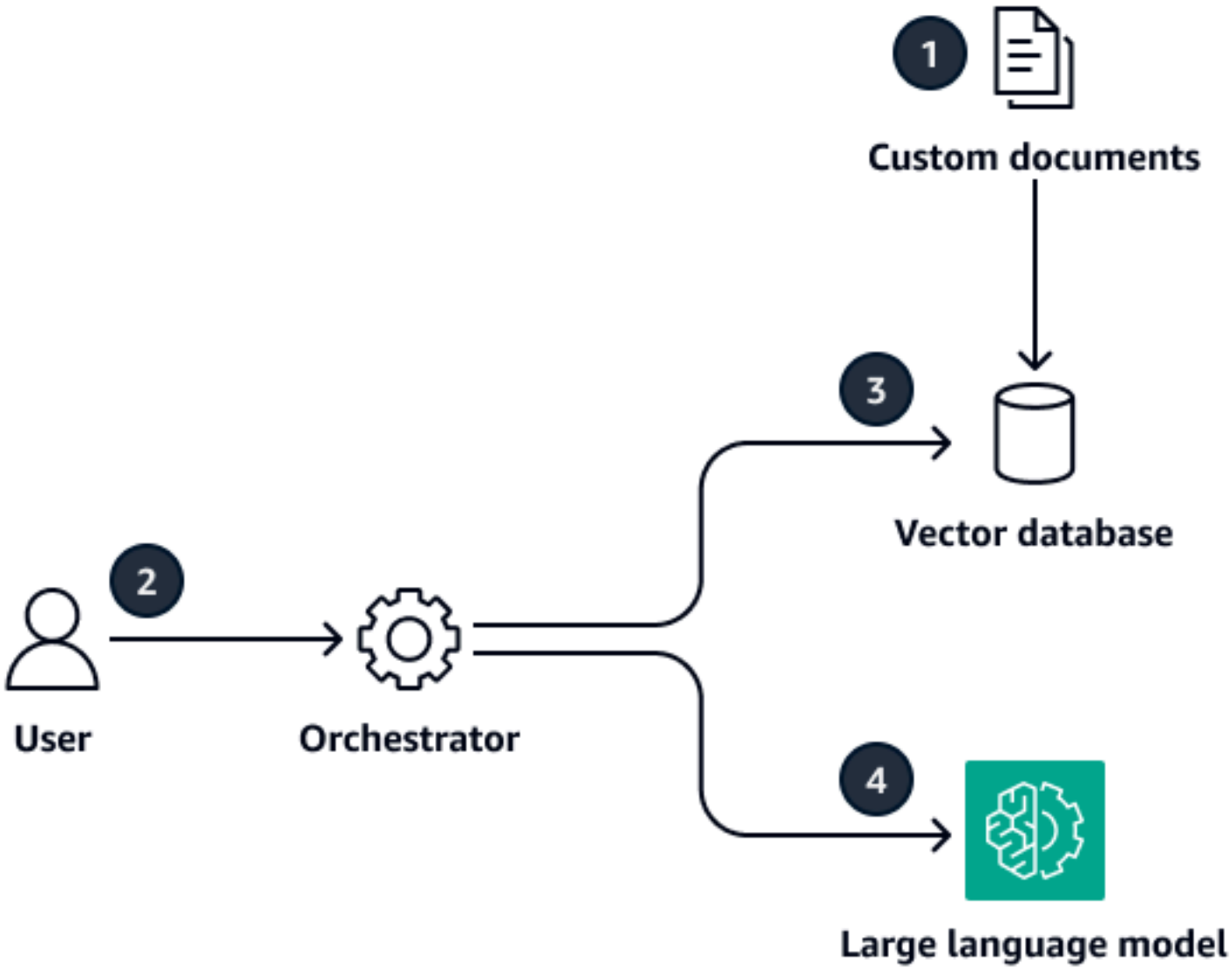
Die letzte Option ist die Verwendung von RAG. Bei RAG referenziert das Foundation-Modell Ihre benutzerdefinierten Dokumente, bevor eine Antwort generiert wird. RAG erweitert die Funktionen des Modells auf die interne Wissensdatenbank Ihres Unternehmens, ohne dass das Modell neu trainiert werden muss. Es ist ein kostengünstiger Ansatz zur Verbesserung der Modellergebnisse, sodass sie in verschiedenen Kontexten relevant, genau und nützlich bleiben.

Themen in diesem Abschnitt:

- Grundlegendes zu Retrieval Augmented Generation
- Vergleich von Retrieval, Augmented Generation und Feinabstimmung
- Anwendungsfälle für Retrieval Augmented Generation

Grundlegendes zu Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) ist eine Technik, mit der ein Large Language Model (LLM) um externe Daten, wie z. B. interne Dokumente eines Unternehmens, erweitert wird. Dadurch erhält das Modell den Kontext, den es benötigt, um genaue und nützliche Ergebnisse für Ihren spezifischen Anwendungsfall zu erzeugen. RAG ist ein pragmatischer und effektiver Ansatz für den Einsatz von LLMs in einem Unternehmen. Das folgende Diagramm zeigt einen allgemeinen Überblick über die Funktionsweise eines RAG-Ansatzes.



Im Großen und Ganzen besteht der RAG-Prozess aus vier Schritten. Der erste Schritt wird einmal ausgeführt, und die anderen drei Schritte werden so oft wie nötig ausgeführt:

1. Sie erstellen Einbettungen, um die internen Dokumente in eine Vektordatenbank aufzunehmen. Einbettungen sind numerische Darstellungen von Text in den Dokumenten, die die semantische oder kontextuelle Bedeutung der Daten erfassen. Eine Vektordatenbank ist im Wesentlichen eine Datenbank dieser Einbettungen und wird manchmal auch als Vektorspeicher oder Vektorindex bezeichnet. Dieser Schritt erfordert das Bereinigen, Formatieren und Aufteilen von Daten, aber dies ist eine einmalige Aktivität, die im Voraus erfolgt.
2. Ein Mensch reicht eine Anfrage in natürlicher Sprache ein.
3. Ein Orchestrator führt eine Ähnlichkeitssuche in der Vektordatenbank durch und ruft die relevanten Daten ab. Der Orchestrator fügt die abgerufenen Daten (auch als Kontext bezeichnet) zur Eingabeaufforderung hinzu, die die Abfrage enthält.
4. Der Orchestrator sendet die Abfrage und den Kontext an das LLM. Das LLM generiert eine Antwort auf die Abfrage, indem es den zusätzlichen Kontext verwendet.

Aus der Sicht eines Benutzers sieht es so aus, als würde RAG mit einem beliebigen LLM interagieren. Das System weiß jedoch viel mehr über die fraglichen Inhalte und bietet Antworten, die genau auf die Wissensbasis des Unternehmens abgestimmt sind.

Weitere Informationen zur Funktionsweise eines RAG-Ansatzes finden Sie auf der AWS Website unter Was ist RAG.

Komponenten von RAG-Systemen auf Produktionsebene

Der Aufbau eines RAG-Systems auf Produktionsebene erfordert das Durchdenken verschiedener Aspekte des RAG-Workflows. Konzeptionell erfordert ein RAG-Workflow auf Produktionsebene unabhängig von der spezifischen Implementierung die folgenden Funktionen und Komponenten:

- Konnektoren — Diese verbinden verschiedene Unternehmensdatenquellen mit der Vektordatenbank. Beispiele für strukturierte Datenquellen sind Transaktions- und Analysedatenbanken. Beispiele für unstrukturierte Datenquellen sind Objektspeicher, Codebasen und SaaS-Plattformen (Software as a Service). Für jede Datenquelle sind möglicherweise unterschiedliche Verbindungsmuster, Lizenzen und Konfigurationen erforderlich.
- Datenverarbeitung — Daten liegen in vielen Formen vor, z. B. als PDFs, gescannte Bilder, Dokumente, Präsentationen und Microsoft SharePoint Dateien. Sie müssen Datenverarbeitungstechniken verwenden, um die Daten zu extrahieren, zu verarbeiten und für die Indexierung vorzubereiten.

- Einbettungen — Um eine Relevanzsuche durchzuführen, müssen Sie Ihre Dokumente und Benutzerabfragen in ein kompatibles Format konvertieren. Mithilfe von eingebetteten Sprachmodellen konvertieren Sie die Dokumente in eine numerische Darstellung. Dies sind im Wesentlichen Eingaben für das zugrunde liegende Fundamentmodell.
- Vektordatenbank — Die Vektordatenbank ist ein Index der Einbettungen, des zugehörigen Textes und der Metadaten. Der Index ist für die Suche und den Abruf optimiert.
- Retriever — Für die Benutzerabfrage ruft der Retriever den relevanten Kontext aus der Vektordatenbank ab und ordnet die Antworten auf der Grundlage der Geschäftsanforderungen.
- Basismodell — Das Basismodell für ein RAG-System ist in der Regel ein LLM. Durch die Verarbeitung des Kontextes und der Aufforderung generiert und formatiert das Foundation-Modell eine Antwort für den Benutzer.
- Leitplanken — Leitplanken sollen sicherstellen, dass die Anfrage, die Aufforderung, der abgerufene Kontext und die LLM-Antwort korrekt, verantwortungsbewusst, ethisch und frei von Halluzinationen und Vorurteilen sind.
- Orchestrator — Der Orchestrator ist für die Planung und Verwaltung des gesamten Workflows verantwortlich.
- Benutzererfahrung — In der Regel interagiert der Benutzer mit einer Konversationsschnittstelle, die über umfangreiche Funktionen verfügt, darunter die Anzeige des Chat-Verlaufs und das Sammeln von Benutzerfeedback zu Antworten.
- Identitäts- und Benutzerverwaltung — Es ist wichtig, den Benutzerzugriff auf die Anwendung genau zu kontrollieren. In der werden Richtlinien AWS Cloud, Rollen und Berechtigungen in der Regel über AWS Identity and Access Management (IAM) verwaltet.

Es liegt auf der Hand, dass die Planung, Entwicklung, Veröffentlichung und Verwaltung eines RAG-Systems mit einem erheblichen Arbeitsaufwand verbunden ist. Vollständig verwaltete Services wie Amazon Bedrock oder Amazon Q Business können Ihnen helfen, einen Teil der undifferenzierten Schwerarbeit zu bewältigen. Benutzerdefinierte RAG-Architekturen können jedoch mehr Kontrolle über die Komponenten wie den Retriever oder die Vektordatenbank bieten.

Vergleich von Retrieval, Augmented Generation und Feinabstimmung

In der folgenden Tabelle werden die Vor- und Nachteile der Feinabstimmung und der Ansätze beschrieben. RAG-based

Ansatz	Vorteile	Nachteile
Fine-tuning	• Wenn ein fein abgestimmtes Modell mithilfe des unbeaufsichtigten Ansatzes trainiert wird, ist es in der Lage, Inhalte zu erstellen, die dem Stil Ihrer Organisation besser entsprechen. • Ein fein abgestimmtes Modell, das auf firmeneigenen oder regulatorischen Daten trainiert wurde, kann Ihrem Unternehmen helfen, interne oder branchenspezifische Daten- und Compliance-Standards einzuhalten.	• Fine-tuning kann je nach Größe des Modells einige Stunden bis Tage dauern. Daher ist es keine gute Lösung, wenn sich Ihre benutzerdefinierten Dokumente häufig ändern. • Fine-tuning erfordert ein Verständnis von Techniken wie Low-Rank Adaptation (LoRa) und parameter effizientes Fine-Tuning (PEFT). Fine-tuning erfordert möglicherweise einen Datenwissenschaftler. • Fine-tuning ist möglicherweise nicht für alle Modelle verfügbar. • Fine-tuned Modelle geben in ihren Antworten keinen Verweis auf die Quelle an. • Bei der Verwendung eines fein abgestimmten Modells zur Beantwortung von Fragen kann das Halluzinationsrisiko erhöht sein.
RAG	• Mit RAG können Sie ohne Feinabstimmung ein System zur Beantwortung von Fragen für Ihre benutzerdefinierten Dokumente erstellen.	• RAG funktioniert nicht gut, wenn es darum geht, Informationen aus ganzen Dokumenten zusammenzufassen.

Ansatz	Vorteile	Nachteile
	• RAG kann die neuesten Dokumente in wenigen Minuten integrieren. • AWS bietet vollständig verwaltete RAG-Lösun gen. Daher sind weder Datenwissenschaftler noch Spezialkenntnisse im Bereich maschinelles Lernen erforderlich. • In seiner Antwort gibt ein RAG-Modell einen Verweis auf die Informationsquelle. • Da RAG den Kontext aus der Vektorsuche als Grundlage für die generierte Antwort verwendet, besteht ein geringeres Halluzina tionsrisiko.	

Wenn Sie eine Lösung zur Beantwortung von Fragen entwickeln müssen, die auf Ihre benutzerdefinierten Dokumente verweist, empfehlen wir Ihnen, von einem Ansatz auszugehen. RAG-based Verwenden Sie die Feinabstimmung, wenn Sie das Modell für zusätzliche Aufgaben, wie z. B. die Zusammenfassung, benötigen.

Sie können die Feinabstimmungs- und RAG-Ansätze in einem einzigen Modell kombinieren. In diesem Fall ändert sich die RAG-Architektur nicht, aber das LLM, das die Antwort generiert, wird ebenfalls an die benutzerdefinierten Dokumente angepasst. Dies kombiniert das Beste aus beiden Welten und könnte eine optimale Lösung für Ihren Anwendungsfall sein. Weitere Informationen zur Kombination von überwachter Feinabstimmung mit RAG finden Sie in der Studie RAFT: Adapting Language Model to Domain Specific RAG von derUniversity of California, Berkeley.

Anwendungsfälle für Retrieval Augmented Generation

Im Folgenden sind gängige Anwendungsfälle für die Verwendung eines RAG-Ansatzes aufgeführt:

- Suchmaschinen — RAG-enabled Suchmaschinen können genauere und aktuellere Featured Snippets in ihren Suchergebnissen bereitstellen.
- Question-answering Systeme — RAG kann die Qualität der Antworten in Systemen zur Beantwortung von Fragen verbessern. Das auf Abrufen basierende Modell verwendet die Ähnlichkeitssuche, um relevante Passagen oder Dokumente zu finden, die die Antwort enthalten. Anschließend generiert es auf der Grundlage dieser Informationen eine präzise und relevante Antwort.
- Einzelhandel oder E-Commerce — RAG kann das Benutzererlebnis im E-Commerce verbessern, indem es relevantere und personalisierte Produktempfehlungen gibt. Durch das Abrufen und Integrieren von Informationen über Benutzerpräferenzen und Produktdetails kann RAG genauere und hilfreichere Empfehlungen für Kunden erstellen.
- Industrie oder Fertigung — In der Fertigung hilft Ihnen RAG dabei, schnell auf wichtige Informationen zuzugreifen, z. B. über den Betrieb von Fabrikanlagen. Es kann auch bei Entscheidungsprozessen, bei der Fehlerbehebung und bei organisatorischen Innovationen helfen. Für Hersteller, die innerhalb strenger regulatorischer Rahmenbedingungen arbeiten, kann RAG schnell aktualisierte Vorschriften und Compliance-Standards aus internen und externen Quellen abrufen, z. B. aus Industriestandards oder Aufsichtsbehörden.
- Gesundheitswesen — RAG hat Potenzial in der Gesundheitsbranche, wo der Zugang zu genauen und aktuellen Informationen von entscheidender Bedeutung ist. Durch das Abrufen und Integrieren von relevantem medizinischem Wissen aus externen Quellen kann RAG genauere und kontextsensivere Antworten in Anwendungen im Gesundheitswesen bereitstellen. Solche Anwendungen erweitern die Informationen, auf die ein menschlicher Arzt zugreifen kann, der letztlich die Entscheidung trifft und nicht das Modell.
- Rechtliches — RAG kann hervorragend in rechtlichen Szenarien wie Fusionen und Übernahmen eingesetzt werden, bei denen komplexe Rechtsdokumente den Kontext für Anfragen bieten. Dies kann Juristen helfen, komplexe regulatorische Probleme schnell zu lösen.

Vollständig verwaltete Optionen für Retrieval Augmented Generation auf AWS

Um Workflows mit Retrieval Augmented Generation (RAG) zu verwalten AWS, können Sie benutzerdefinierte RAG-Pipelines verwenden oder einige der Funktionen der vollständig verwalteten Dienste nutzen, die die Lösung bietet. AWS Da sie viele der Kernkomponenten eines RAG-based Systems enthalten, können Fully Managed Services Ihnen helfen, einen Teil der undifferenzierten Schwerarbeit zu bewältigen. Diese Dienste bieten jedoch weniger Möglichkeiten zur Anpassung.

Die vollständig verwalteten AWS-Services Systeme verwenden Konnektoren, um Daten aus externen Datenquellen wie Websites, Atlassian Confluence oder Microsoft aufzunehmen. SharePoint Die unterstützten Datenquellen variieren je nach. AWS-Service

In diesem Abschnitt werden die folgenden vollständig verwalteten Optionen für die Erstellung von RAG-Workflows untersucht AWS:

- Wissensdatenbanken für Amazon Bedrock
- Amazon Q Business
- Amazon SageMaker AI-Leinwand

Weitere Informationen darüber, wie Sie zwischen diesen Optionen wählen können, finden Sie Wählen Sie eine Option zum Abrufen erweiterter Generierung auf AWS in diesem Handbuch.

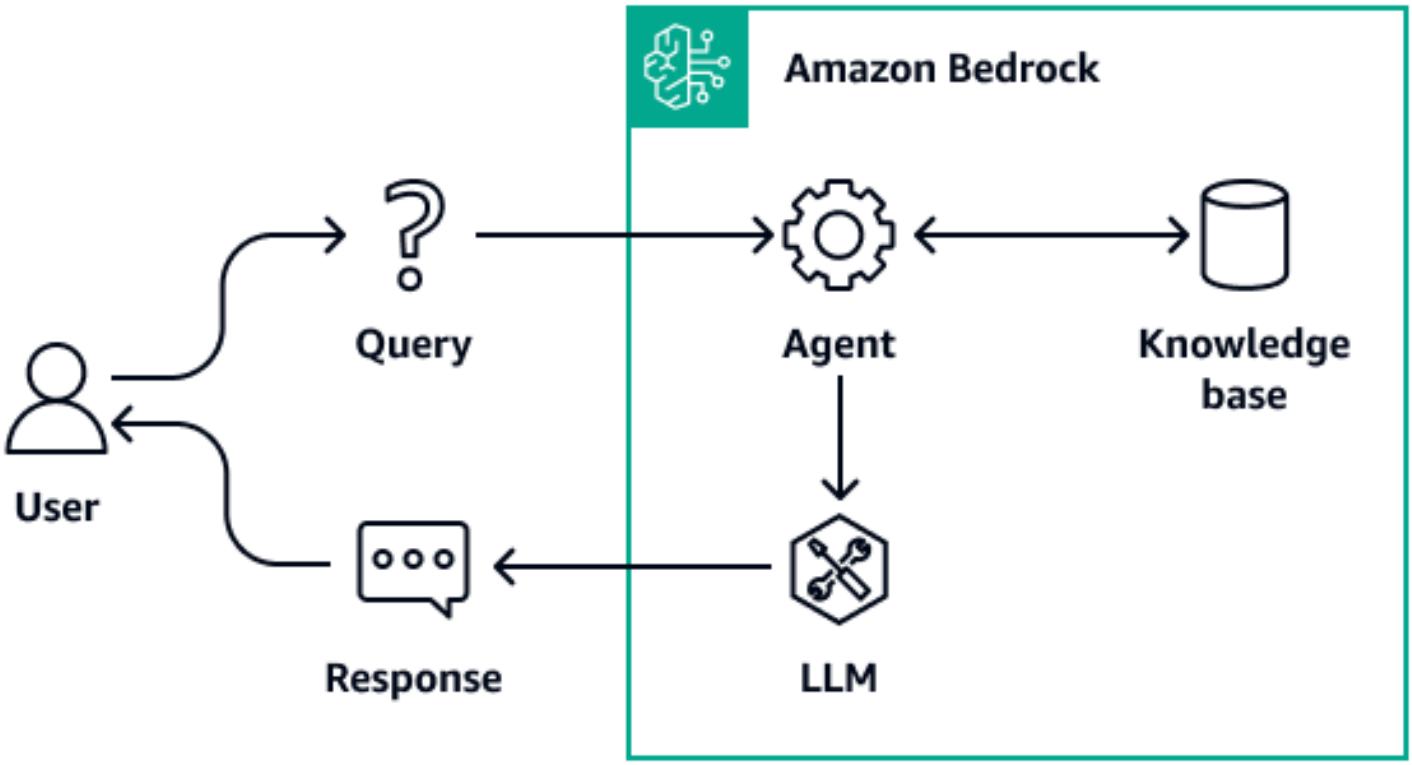
Wissensdatenbanken für Amazon Bedrock

Amazon Bedrock ist ein vollständig verwalteter Service, der Ihnen leistungsstarke Foundation-Modelle (FMs) von führenden KI-Startups und Amazon über eine einheitliche API zur Verfügung stellt. Knowledge Bases ist eine Funktion von Amazon Bedrock, mit der Sie den gesamten RAG-Workflow implementieren können, von der Aufnahme über den Abruf bis hin zur sofortigen Erweiterung. Es ist nicht erforderlich, benutzerdefinierte Integrationen für Datenquellen zu erstellen oder Datenflüsse zu verwalten. Das Sitzungskontextmanagement ist integriert, sodass Ihre generative KI-Anwendung problemlos Multi-Turn-Konversationen unterstützen kann.

Nachdem Sie den Speicherort Ihrer Daten angegeben haben, ruft Knowledge Bases für Amazon Bedrock die Dokumente intern ab, teilt sie in Textblöcke auf, konvertiert den Text in Einbettungen

und speichert die Einbettungen dann in der Vektordatenbank Ihrer Wahl. Amazon Bedrock verwaltet und aktualisiert die Einbettungen und sorgt dafür, dass die Vektordatenbank mit den Daten synchron bleibt. Weitere Informationen zur Funktionsweise von Wissensdatenbanken finden Sie unter So funktionieren Amazon Bedrock-Wissensdatenbanken.

Wenn Sie einem Amazon Bedrock-Agenten Wissensdatenbanken hinzufügen, identifiziert der Agent die entsprechende Wissensdatenbank auf der Grundlage der Benutzereingabe. Der Agent ruft die relevanten Informationen ab und fügt sie der Eingabeaufforderung hinzu. Die aktualisierte Aufforderung stellt dem Modell mehr Kontextinformationen zur Verfügung, um eine Antwort zu generieren. Um die Transparenz zu erhöhen und Halluzinationen zu minimieren, können die aus der Wissensdatenbank abgerufenen Informationen bis zu ihrer Quelle zurückverfolgt werden.



Amazon Bedrock unterstützt die folgenden zwei APIs für RAG:

- RetrieveAndGenerate— Sie können diese API verwenden, um Ihre Wissensdatenbank abzufragen und anhand der abgerufenen Informationen Antworten zu generieren. Intern wandelt Amazon Bedrock die Abfragen in Einbettungen um, fragt die Wissensdatenbank ab, ergänzt die Eingabeaufforderung mit den Suchergebnissen als Kontextinformationen und gibt die Antwort zurück. LLM-generated Amazon Bedrock verwaltet auch das Kurzzeitgedächtnis der Konversation, um kontextbezogenere Ergebnisse zu erzielen.

- Abrufen — Sie können diese API verwenden, um Ihre Wissensdatenbank mit Informationen abzufragen, die direkt aus der Wissensdatenbank abgerufen wurden. Sie können die von dieser API zurückgegebenen Informationen verwenden, um den abgerufenen Text zu verarbeiten, seine Relevanz zu bewerten oder einen separaten Workflow für die Antwortgenerierung zu entwickeln. Intern wandelt Amazon Bedrock die Abfragen in Einbettungen um, durchsucht die Wissensdatenbank und gibt die entsprechenden Ergebnisse zurück. Sie können zusätzliche Workflows zusätzlich zu den Suchergebnissen erstellen. Sie können das LangChain`AmazonKnowledgeBasesRetriever`Plugin beispielsweise verwenden, um RAG-Workflows in generative KI-Anwendungen zu integrieren.

Architekturmuster und schrittweise Anleitungen zur Verwendung der APIs finden Sie unter Knowledge Bases now provides fully managed RAG experience in Amazon Bedrock (AWS Blogbeitrag). Weitere Informationen zur Verwendung der `RetrieveAndGenerate` API zum Erstellen eines RAG-Workflows für eine intelligente Chat-basierte Anwendung finden Sie unter Erstellen einer kontextbezogenen Chatbot-Anwendung mithilfe von Amazon Bedrock Knowledge Bases (Blogbeitrag).AWS

Datenquellen für Wissensdatenbanken

Sie können Ihre eigenen Daten mit einer Wissensdatenbank verbinden. Nachdem Sie einen Datenquellen-Connector konfiguriert haben, können Sie Ihre Daten mit Ihrer Wissensdatenbank synchronisieren oder auf dem neuesten Stand halten und Ihre Daten für Abfragen zur Verfügung stellen. Amazon Bedrock Knowledge Bases unterstützen Verbindungen zu den folgenden Datenquellen:

- Amazon Simple Storage Service (Amazon S3) — Sie können einen Amazon S3 S3-Bucket mit einer Amazon Bedrock-Wissensdatenbank verbinden, indem Sie entweder die Konsole oder die API verwenden. Die Wissensdatenbank nimmt die Dateien im Bucket auf und indexiert sie. Diese Art von Datenquelle unterstützt die folgenden Funktionen:
  - Metadatenfelder für Dokumente — Sie können eine separate Datei hinzufügen, um die Metadaten für die Dateien im Amazon S3 S3-Bucket anzugeben. Sie können diese Metadatenfelder dann verwenden, um die Relevanz von Antworten zu filtern und zu verbessern.
  - Inklusions- oder Ausschlussfilter — Sie können beim Crawlen bestimmte Inhalte ein- oder ausschließen.
  - Inkrementelle Synchronisierung — Die Inhaltsänderungen werden nachverfolgt, und es werden nur Inhalte gecrawlt, die sich seit der letzten Synchronisierung geändert haben.

- Confluence— Sie können eine Atlassian Confluence Instance mithilfe der Konsole oder der API mit einer Amazon Bedrock-Wissensdatenbank verbinden. Diese Art von Datenquelle unterstützt die folgenden Funktionen:
  - Automatische Erkennung der wichtigsten Dokumentfelder — Die Metadatenfelder werden automatisch erkannt und gecrawlt. Sie können diese Felder zum Filtern verwenden.
  - Inhaltsfilter zum Ein- oder Ausschließen — Sie können bestimmte Inhalte ein- oder ausschließen, indem Sie ein Präfix oder ein Muster mit regulären Ausdrücken für den Bereich, den Seitentitel, den Blogtitel, den Kommentar, den Namen des Anhangs oder die Erweiterung verwenden.
  - Inkrementelle Synchronisierung — Die Inhaltsänderungen werden nachverfolgt, und es werden nur Inhalte gecrawlt, die sich seit der letzten Synchronisierung geändert haben.
  - OAuth 2.0-Authentifizierung, Authentifizierung mit Confluence API-Token — Die Authentifizierungsdaten werden in gespeichert. AWS Secrets Manager
- Microsoft SharePoint— Sie können eine SharePoint Instanz mit einer Wissensdatenbank verbinden, indem Sie entweder die Konsole oder die API verwenden. Diese Art von Datenquelle unterstützt die folgenden Funktionen:
  - Automatische Erkennung der wichtigsten Dokumentfelder — Die Metadatenfelder werden automatisch erkannt und gecrawlt. Sie können diese Felder zum Filtern verwenden.
  - Inhaltsfilter zum Ein- oder Ausschließen — Sie können bestimmte Inhalte ein- oder ausschließen, indem Sie ein Präfix oder ein Muster mit regulären Ausdrücken für den Titel der Hauptseite, den Veranstaltungsnamen und den Dateinamen (einschließlich der Erweiterung) verwenden.
  - Inkrementelle Synchronisierung — Die Inhaltsänderungen werden nachverfolgt, und es werden nur Inhalte gecrawlt, die sich seit der letzten Synchronisierung geändert haben.
  - OAuth 2.0-Authentifizierung — Die Authentifizierungsdaten werden in gespeichert. AWS Secrets Manager
- Salesforce— Sie können eine Salesforce Instanz mit einer Wissensdatenbank verbinden, indem Sie entweder die Konsole oder die API verwenden. Diese Art von Datenquelle unterstützt die folgenden Funktionen:
  - Automatische Erkennung der wichtigsten Dokumentfelder — Die Metadatenfelder werden automatisch erkannt und gecrawlt. Sie können diese Felder zum Filtern verwenden.
  - Inhaltsfilter zum Einschließen oder Ausschließen — Sie können bestimmte Inhalte mit einem Präfix oder einem Muster für reguläre Ausdrücke ein- oder ausschließen. Eine Liste der Inhaltstypen, auf die Sie Filter anwenden können, finden Sie unter Inclusion/exclusion Filter in der Amazon Bedrock-Dokumentation.

  - Inkrementelle Synchronisierung — Die Inhaltsänderungen werden nachverfolgt, und es werden nur Inhalte gecrawlt, die sich seit der letzten Synchronisierung geändert haben.
  - OAuth 2.0-Authentifizierung — Die Authentifizierungsdaten werden in gespeichert. AWS Secrets Manager
- Webcrawler — Ein Amazon Bedrock Web Crawler stellt eine Verbindung zu den von Ihnen angegebenen URLs her und crawlt sie. Die folgenden Funktionen werden unterstützt:
  - Wählen Sie mehrere URLs zum Crawlen aus
  - Beachten Sie die Standardanweisungen von robots.txt wie und `Allow` `Disallow`
  - Schließt URLs aus, die einem Muster entsprechen
  - Beschränken Sie die Crawling-Rate
  - Sehen Sie CloudWatch sich in Amazon den Status jeder gecrawlten URL an

Weitere Informationen zu den Datenquellen, die Sie mit Ihrer Amazon Bedrock-Wissensdatenbank verbinden können, finden Sie unter Erstellen eines Datenquellen-Connectors für Ihre Wissensdatenbank.

Vektor-Datenbanken für Wissensdatenbanken

Wenn Sie eine Verbindung zwischen der Wissensdatenbank und der Datenquelle einrichten, müssen Sie eine Vektordatenbank konfigurieren, die auch als Vektorspeicher bezeichnet wird. In einer Vektordatenbank speichert, aktualisiert und verwaltet Amazon Bedrock die Einbettungen, die Ihre Daten darstellen. Jede Datenquelle unterstützt verschiedene Arten von Vektordatenbanken. Informationen darüber, welche Vektordatenbanken für Ihre Datenquelle verfügbar sind, finden Sie unter Datenquellentypen.

Wenn Sie es vorziehen, dass Amazon Bedrock automatisch eine Vektordatenbank in Amazon OpenSearch Serverless für Sie erstellt, können Sie diese Option bei der Erstellung der Wissensdatenbank wählen. Sie können sich jedoch auch dafür entscheiden, Ihre eigene Vektordatenbank einzurichten. Wenn Sie Ihre eigene Vektordatenbank einrichten, finden Sie eine Wissensdatenbank unter Voraussetzungen für Ihren eigenen Vektorspeicher. Jeder Vektordatenbanktyp hat seine eigenen Voraussetzungen.

Abhängig von Ihrem Datenquellentyp unterstützen die Amazon Bedrock-Wissensdatenbanken die folgenden Vektordatenbanken:

- Amazon OpenSearch Serverlos

- Amazon Aurora PostgreSQL-Compatible Aurora-Ausgabe
- Pinecone(PineconeDokumentation)
- Redis Enterprise Cloud(RedisDokumentation)
- MongoDB Atlas(MongoDBDokumentation)

Amazon Q Business

Amazon Q Business ist ein vollständig verwalteter, auf generativer KI basierender Assistent, den Sie so konfigurieren können, dass er Fragen beantwortet, Zusammenfassungen bereitstellt, Inhalte generiert und Aufgaben auf der Grundlage Ihrer Unternehmensdaten erledigt. Es ermöglicht Endbenutzern, sofortige, berechtigungsabhängige Antworten aus Unternehmensdatenquellen mit Quellenangaben zu erhalten.

Schlüssel-Features

Die folgenden Funktionen von Amazon Q Business können Ihnen bei der Entwicklung einer RAG-based generativen KI-Anwendung in Produktionsqualität helfen:

- Built-in Konnektoren — Amazon Q Business unterstützt mehr als 40 Arten von Konnektoren, z. B. Konnektoren für Adobe Experience Manager (AEM)Salesforce,Jira, undMicrosoft SharePoint. Eine vollständige Liste finden Sie unter Unterstützte Konnektoren. Wenn Sie einen Connector benötigen, der nicht unterstützt wird, können Sie Amazon verwenden, AppFlow um Daten aus Ihrer Datenquelle in Amazon Simple Storage Service (Amazon S3) abzurufen und dann Amazon Q Business mit dem Amazon S3-Bucket zu verbinden. Eine vollständige Liste der Datenquellen, die Amazon AppFlow unterstützt, finden Sie unter Unterstützte Anwendungen.
- Built-in Indizierungspipelines — Amazon Q Business bietet eine integrierte Pipeline für die Indizierung von Daten in einer Vektordatenbank. Sie können eine AWS Lambda Funktion verwenden, um Vorverarbeitungslogik für Ihre Indexierungspipeline hinzuzufügen.
- Indexoptionen — Sie können einen systemeigenen Index in Amazon Q Business erstellen und bereitstellen, und Sie verwenden einen Amazon Q Business Retriever, um Daten aus diesem Index abzurufen. Alternativ können Sie einen vorkonfigurierten Amazon Kendra Kendra-Index als Retriever verwenden. Weitere Informationen finden Sie unter Einen Retriever für eine Amazon Q Business-Anwendung erstellen.

- Foundation-Modelle — Amazon Q Business verwendet die Foundation-Modelle, die in Amazon Bedrock unterstützt werden. Eine vollständige Liste finden Sie unter Unterstützte Foundation-Modelle in Amazon Bedrock.
- Plugins — Amazon Q Business bietet die Möglichkeit, Plugins zur Integration in Zielsysteme zu verwenden, z. B. eine automatisierte Methode zur Zusammenfassung von Ticketinformationen und zur Ticketerstellung inJira. Nach der Konfiguration können Plugins Lese- und Schreibaktionen unterstützen, die Ihnen helfen können, die Produktivität der Endbenutzer zu steigern. Amazon Q Business unterstützt zwei Arten von Plug-ins: integrierte Plug-ins und benutzerdefinierte Plug-ins.
- Guardrails — Amazon Q Business unterstützt globale Kontrollen und Kontrollen auf Themenebene. Diese Kontrollen können beispielsweise personenbezogene Daten (PII), Missbrauch oder vertrauliche Informationen in Eingabeaufforderungen erkennen. Weitere Informationen finden Sie unter Administratorkontrollen und Leitplanken in Amazon Q Business.
- Identitätsmanagement — Mit Amazon Q Business können Sie Benutzer und ihren Zugriff auf die RAG-based generative KI-Anwendung verwalten. Weitere Informationen finden Sie unter Identitäts- und Zugriffsverwaltung für Amazon Q Business. Außerdem indexieren Amazon Q Business Connectors Informationen zur Zugriffskontrollliste (ACL), die zusammen mit dem Dokument selbst an ein Dokument angehängt sind. Anschließend speichert Amazon Q Business die von ihm indizierten ACL-Informationen im Amazon Q Business User Store, um Benutzer- und Gruppenzuordnungen zu erstellen und Chat-Antworten basierend auf dem Zugriff des Endbenutzers auf Dokumente zu filtern. Weitere Informationen finden Sie unter Konzepte für Datenquellenkonnektoren.
- Anreicherung von Dokumenten — Mit der Funktion zur Anreicherung von Dokumenten können Sie kontrollieren, welche Dokumente und Dokumentattribute in Ihren Index aufgenommen werden und wie sie aufgenommen werden. Dies kann auf zwei Arten erreicht werden:
  - Grundoperationen konfigurieren — Verwenden Sie grundlegende Operationen, um Dokumentattribute zu Ihren Daten hinzuzufügen, zu aktualisieren oder zu löschen. Sie können beispielsweise personenbezogene Daten löschen, indem Sie alle Dokumentattribute löschen, die sich auf personenbezogene Daten beziehen.
  - Lambda-Funktionen konfigurieren — Verwenden Sie eine vorkonfigurierte Lambda-Funktion, um eine individuellere, erweiterte Logik zur Manipulation von Dokumentattributen auf Ihre Daten anzuwenden. Beispielsweise können Ihre Unternehmensdaten als gescannte Bilder gespeichert werden. In diesem Fall können Sie eine Lambda-Funktion verwenden, um die optische Zeichenerkennung (OCR) für die gescannten Dokumente auszuführen, um Text aus ihnen zu extrahieren. Anschließend wird jedes gescannte Dokument bei der Aufnahme als

Textdokument behandelt. Schließlich berücksichtigt Amazon Q während des Chats die aus den gescannten Dokumenten extrahierten Textdaten, wenn es Antworten generiert.

Bei der Implementierung Ihrer Lösung können Sie wählen, ob Sie beide Ansätze zur Anreicherung von Dokumenten kombinieren möchten. Sie können grundlegende Operationen verwenden, um eine erste Analyse Ihrer Daten durchzuführen, und dann eine Lambda-Funktion für komplexere Operationen verwenden. Weitere Informationen finden Sie unter Anreicherung von Dokumenten in Amazon Q Business.

- Integration — Nachdem Sie Ihre Amazon Q Business-Anwendung erstellt haben, können Sie sie in andere Anwendungen integrieren, z. B. Slack oderMicrosoft Teams. Siehe beispielsweise Bereitstellen eines Slack Gateways für Amazon Q Business und Bereitstellen eines Microsoft Teams Gateways für Amazon Q Business (AWS Blogbeiträge).

End-user Anpassung

Amazon Q Business unterstützt das Hochladen von Dokumenten, die möglicherweise nicht in den Datenquellen und im Index Ihrer Organisation gespeichert sind. Hochgeladene Dokumente werden nicht gespeichert. Sie können nur für die Konversation verwendet werden, in der die Dokumente hochgeladen werden. Amazon Q Business unterstützt bestimmte Dokumenttypen für das Hochladen. Weitere Informationen finden Sie unter Dateien hochladen und chatten in Amazon Q Business.

Amazon Q Business beinhaltet eine Funktion zum Filtern nach Dokumentattributen. Sowohl Administratoren als auch Endbenutzer können diese Funktion verwenden. Administratoren können die Chat-Antworten für Endbenutzer mithilfe von Attributen anpassen und steuern. Wenn es sich bei dem Datenquellentyp beispielsweise um ein Attribut handelt, das an Ihre Dokumente angehängt ist, können Sie angeben, dass Chat-Antworten nur aus einer bestimmten Datenquelle generiert werden. Oder Sie können Endbenutzern ermöglichen, den Umfang der Chat-Antworten einzuschränken, indem Sie die von Ihnen ausgewählten Attributfilter verwenden.

Endbenutzer können schlanke, speziell entwickelte Amazon Q Apps in Ihrer breiteren Amazon Q Business-Anwendungsumgebung erstellen. Amazon Q-Apps ermöglichen die Automatisierung von Aufgaben für eine bestimmte Domain, z. B. eine speziell für das Marketingteam entwickelte App.

Amazon SageMaker AI-Leinwand

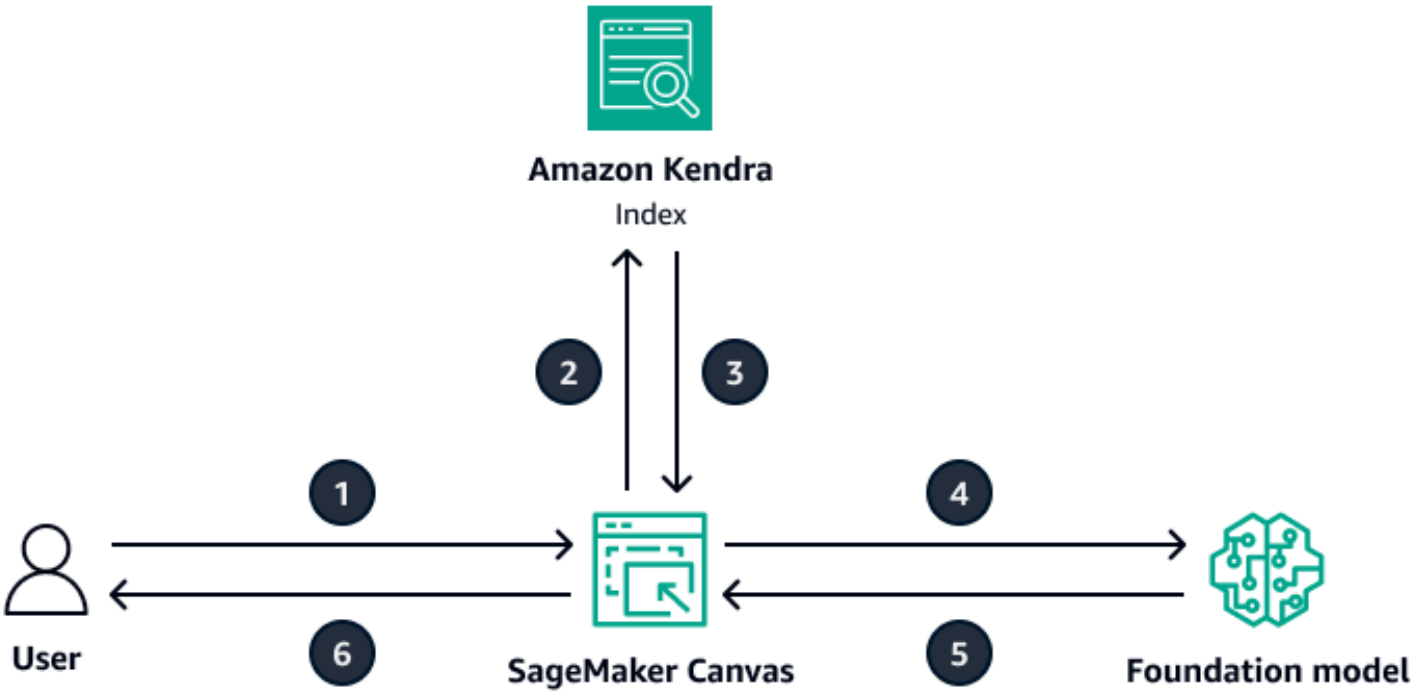
Amazon SageMaker AI Canvas hilft Ihnen, mithilfe von maschinellem Lernen Vorhersagen zu generieren, ohne Code schreiben zu müssen. Es bietet eine visuelle Oberfläche ohne Code,

mit der Sie Daten vorbereiten, ML-Modelle erstellen und bereitstellen und so den gesamten ML-Lebenszyklus in einer einheitlichen Umgebung optimieren können. Die Komplexität von Datenaufbereitung, Modellentwicklung, Erkennung von Verzerrungen, Erklärbarkeit und Überwachung wird hinter einer intuitiven Oberfläche zusammengefasst. Benutzer müssen keine Experten für SageMaker KI oder maschinelles Lernen (MLOps) sein, um Modelle mit AI Canvas zu entwickeln, zu operationalisieren und zu überwachen. SageMaker

Bei SageMaker AI Canvas wird die RAG-Funktionalität über eine Funktion zur Dokumentenabfrage ohne Code bereitgestellt. Sie können das Chat-Erlebnis in SageMaker AI Canvas bereichern, indem Sie einen Amazon Kendra Kendra-Index als zugrunde liegende Unternehmenssuche verwenden. Weitere Informationen finden Sie unter Extrahieren von Informationen aus Dokumenten mit Dokumentenabfrage.

Die Verbindung von SageMaker AI Canvas mit dem Amazon Kendra Kendra-Index erfordert eine einmalige Einrichtung. Im Rahmen der Domain-Konfiguration kann ein Cloud-Administrator einen oder mehrere Kendra-Indizes auswählen, die der Benutzer bei der Interaktion mit SageMaker Canvas abfragen kann. Anweisungen zur Aktivierung der Dokumentabfragefunktion finden Sie unter Erste Schritte mit der Verwendung von Amazon SageMaker AI Canvas.

SageMaker AI Canvas verwaltet die zugrunde liegende Kommunikation zwischen Amazon Kendra und dem ausgewählten Foundation-Modell. Weitere Informationen zu den Basismodellen, die SageMaker AI Canvas unterstützt, finden Sie unter Generative KI-Grundmodelle in SageMaker AI Canvas. Das folgende Diagramm zeigt, wie die Funktion zur Dokumentenabfrage funktioniert, nachdem der Cloud-Administrator SageMaker AI Canvas mit einem Amazon Kendra Kendra-Index verbunden hat.



Das Diagramm zeigt den folgenden Workflow:

1. Der Benutzer startet einen neuen Chat in SageMaker AI Canvas, aktiviert Query documents, wählt den Zielindex aus und reicht dann eine Frage ein.
2. SageMaker AI Canvas verwendet die Abfrage, um den Amazon Kendra Kendra-Index nach relevanten Daten zu durchsuchen.
3. SageMaker AI Canvas ruft die Daten und ihre Quellen aus dem Amazon Kendra Kendra-Index ab.
4. SageMaker AI Canvas aktualisiert die Aufforderung, um den abgerufenen Kontext aus dem Amazon Kendra Kendra-Index aufzunehmen, und leitet die Aufforderung an das Foundation-Modell weiter.
5. Das Foundation-Modell verwendet die ursprüngliche Frage und den abgerufenen Kontext, um eine Antwort zu generieren.
6. SageMaker AI Canvas stellt dem Benutzer die generierte Antwort zur Verfügung. Es enthält Verweise auf die Datenquellen, z. B. Dokumente, die zur Generierung der Antwort verwendet wurden.

Custom Retrieval Augmented Generation-Architekturen auf AWS

Im vorherigen Abschnitt wird beschrieben, wie Sie eine vollständig verwaltete Augmented Generation (RAG) AWS-Service für Retrieval verwenden. In einigen Anwendungsfällen ist jedoch mehr Kontrolle über die Systemkomponenten wie den Retriever oder das LLM (auch Generator genannt) erforderlich. Beispielsweise benötigen Sie möglicherweise die Flexibilität, Ihre eigene Vektordatenbank auszuwählen oder auf eine nicht unterstützte Datenquelle zuzugreifen. Für diese Anwendungsfälle können Sie eine benutzerdefinierte RAG-Architektur erstellen.

In diesem Abschnitt werden folgende Themen behandelt:

- Retriever für RAG-Workflows
- Generatoren für RAG-Workflows

Weitere Informationen zur Auswahl zwischen den Optionen Retriever und Generator in diesem Abschnitt finden Sie Wählen Sie eine Option zum Abrufen erweiterter Generierung auf AWS in diesem Handbuch.

Retriever für RAG-Workflows

In diesem Abschnitt wird erklärt, wie Sie einen Retriever erstellen. Sie können eine vollständig verwaltete semantische Suchlösung wie Amazon Kendra verwenden oder mithilfe einer Vektordatenbank eine benutzerdefinierte semantische Suche erstellen. AWS

Bevor Sie sich mit den Retrieveroptionen befassen, stellen Sie sicher, dass Sie die drei Schritte des Vektorsuchprozesses verstanden haben:

1. Sie teilen die Dokumente, die indexiert werden müssen, in kleinere Teile auf. Dies wird als Chunking bezeichnet.
2. Sie verwenden einen Prozess namens Einbetten, um jeden Chunk in einen mathematischen Vektor umzuwandeln. Anschließend indizieren Sie jeden Vektor in einer Vektordatenbank. Der Ansatz, mit dem Sie die Dokumente indizieren, beeinflusst die Geschwindigkeit und Genauigkeit der Suche. Der Indizierungsansatz hängt von der Vektordatenbank und den von ihr bereitgestellten Konfigurationsoptionen ab.

3. Sie konvertieren die Benutzerabfrage mit demselben Verfahren in einen Vektor. Der Retriever durchsucht die Vektordatenbank nach Vektoren, die dem Abfragevektor des Benutzers ähnlich sind. Die Ähnlichkeit wird anhand von Metriken wie der euklidischen Distanz, der Kosinusdistanz oder dem Punktprodukt berechnet.

In diesem Handbuch wird beschrieben, wie Sie mit den folgenden Diensten AWS-Services oder Diensten von Drittanbietern eine benutzerdefinierte Abruf-Ebene erstellen können: AWS

- Amazon Kendra
- OpenSearch Amazon-Dienst
- Amazon Aurora PostgreSQL und pgvector
- Amazon Neptune Analytics
- Amazon MemoryDB
- Amazon DocumentDB
- Tannenzapfen
- MongoDB-Atlas
- Weben

Amazon Kendra

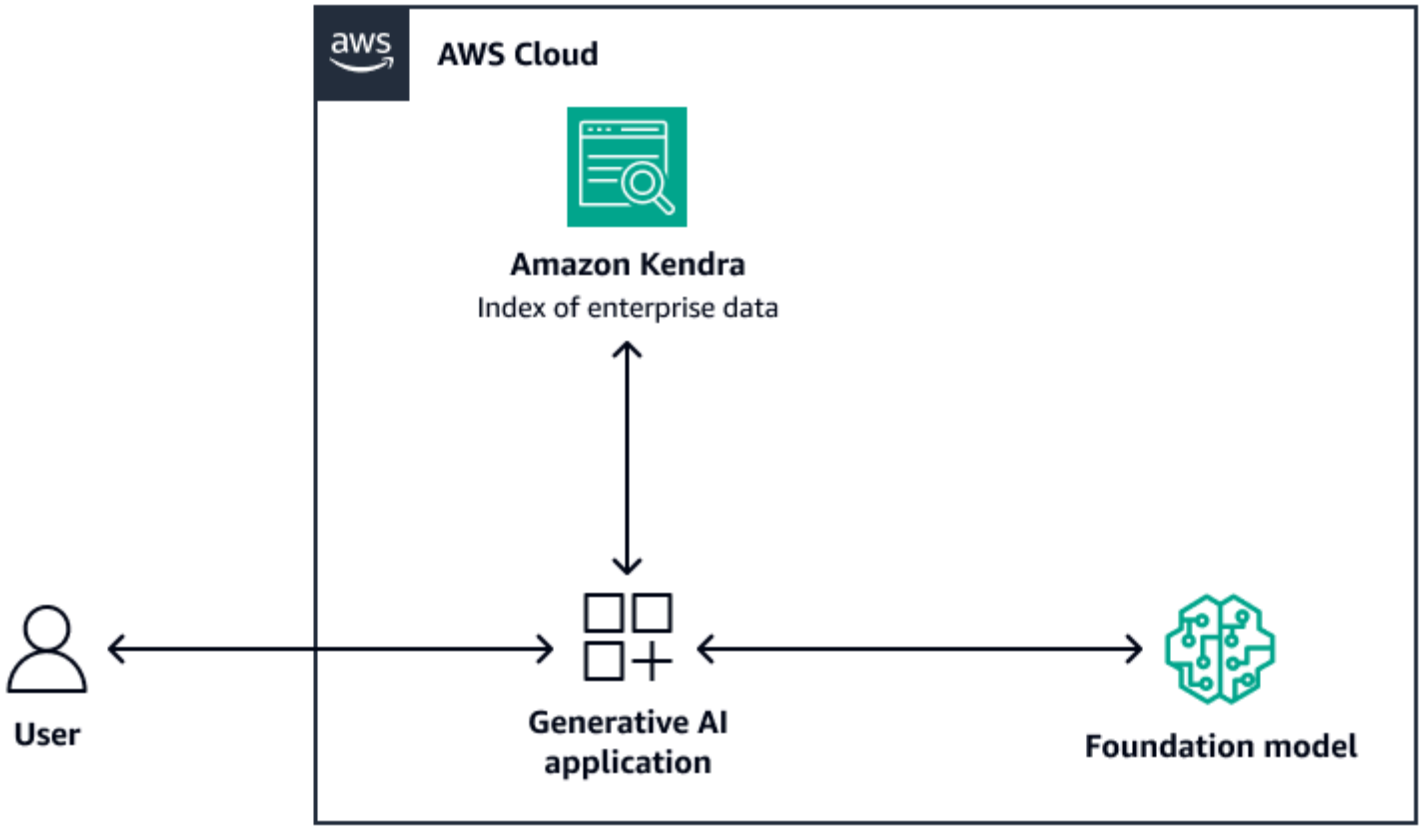
Amazon Kendra ist ein vollständig verwalteter, intelligenter Suchservice, der natürliche Sprachverarbeitung und fortschrittliche Algorithmen für maschinelles Lernen verwendet, um spezifische Antworten auf Suchfragen aus Ihren Daten zurückzugeben. Amazon Kendra hilft Ihnen dabei, Dokumente aus mehreren Quellen direkt aufzunehmen und die Dokumente abzufragen, nachdem sie erfolgreich synchronisiert wurden. Durch den Synchronisierungsprozess wird die erforderliche Infrastruktur geschaffen, um eine Vektorsuche im aufgenommenen Dokument zu erstellen. Daher benötigt Amazon Kendra nicht die traditionellen drei Schritte des Vektorsuchprozesses. Nach der ersten Synchronisierung können Sie einen definierten Zeitplan verwenden, um die laufende Datenaufnahme zu handhaben.

Im Folgenden sind die Vorteile der Verwendung von Amazon Kendra for RAG aufgeführt:

- Sie müssen keine Vektordatenbank verwalten, da Amazon Kendra den gesamten Vektorsuchprozess abwickelt.

- Amazon Kendra enthält vorgefertigte Konnektoren für beliebte Datenquellen wie Datenbanken, Website-Crawler, Amazon S3 S3-Buckets, Microsoft SharePoint Instances und Instances. Atlassian Confluence Von AWS Partnern entwickelte Konnektoren sind verfügbar, z. B. Konnektoren für und. Box GitLab
- Amazon Kendra bietet eine ACL-Filterung (Access Control List), die nur Dokumente zurückgibt, auf die der Endbenutzer Zugriff hat.
- Amazon Kendra kann Antworten auf der Grundlage von Metadaten wie Datum oder Quell-Repository beschleunigen.

Die folgende Abbildung zeigt eine Beispielarchitektur, die Amazon Kendra als Abruf-Ebene des RAG-Systems verwendet. Weitere Informationen finden Sie unter Schnelles Erstellen hochgenauer generativer KI-Anwendungen auf Unternehmensdaten mithilfe von Amazon Kendra und großen Sprachmodellen (AWS Blogbeitrag). LangChain



Für das Foundation-Modell können Sie Amazon Bedrock oder ein über Amazon SageMaker AI bereitgestelltes LLM verwenden. JumpStart Sie können AWS Lambda with verwenden LangChain, um den Fluss zwischen dem Benutzer, Amazon Kendra und dem LLM zu orchestrieren.

Informationen zum Erstellen eines RAG-Systems, das Amazon Kendra und verschiedene LLMs verwendetLangChain, finden Sie im Amazon Kendra LangChain Extensions Repository. GitHub

OpenSearch Amazon-Dienst

Amazon OpenSearch Service bietet integrierte ML-Algorithmen für die Suche nach k-Nearest Neighbours (k-NN), um eine Vektorsuche durchzuführen. OpenSearch Der Service bietet auch eine Vektor-Engine für Amazon EMR Serverless. Mit dieser Vektor-Engine können Sie ein RAG-System erstellen, das über skalierbare und leistungsstarke Vektorspeicher- und Suchfunktionen verfügt. Weitere Informationen zum Erstellen eines RAG-Systems mithilfe von OpenSearch Serverless finden Sie unter Erstellen skalierbarer und serverloser RAG-Workflows mit einer Vektor-Engine für Amazon OpenSearch Serverless- und Amazon Bedrock Claude-Modelle (AWS Blogbeitrag).

Im Folgenden sind die Vorteile der Verwendung von OpenSearch Service für die Vektorsuche aufgeführt:

- Es bietet die vollständige Kontrolle über die Vektordatenbank, einschließlich der Erstellung einer skalierbaren Vektorsuche mithilfe von OpenSearch Serverless.
- Es bietet die Kontrolle über die Chunking-Strategie.
- Es verwendet ANN-Algorithmen (Approximate Nearest Neighbor) aus den Bibliotheken Non-Metric Space Library (NMSLIB), Faiss und Apache Lucene, um eine k-NN-Suche durchzuführen. Sie können den Algorithmus je nach Anwendungsfall ändern. Weitere Informationen zu den Optionen für die Anpassung der Vektorsuche über OpenSearch Service finden Sie unter Erläuterung der Funktionen der Amazon OpenSearch Service-Vektordatenbank (AWS Blogbeitrag).
- OpenSearch Serverless lässt sich als Vektorindex in die Wissensdatenbanken von Amazon Bedrock integrieren.

Amazon Aurora PostgreSQL und pgvector

Amazon Aurora PostgreSQL-Compatible Edition ist eine vollständig verwaltete relationale Datenbank-Engine, die Sie bei der Einrichtung, dem Betrieb und der Skalierung von PostgreSQL-Bereitstellungen unterstützt. pgvector ist eine Open-Source-PostgreSQL-Erweiterung, die Funktionen zur Suche nach Vektorähnlichkeit bietet. Diese Erweiterung ist sowohl für Aurora PostgreSQL-Compatible als auch für Amazon Relational Database Service (Amazon RDS) für PostgreSQL verfügbar. Weitere Informationen zum Aufbau eines RAG-based Systems, das Aurora PostgreSQL-Compatible und pgvector verwendet, finden Sie in den folgenden AWS Blogbeiträgen:

- Aufbau der AI-powered Suche in PostgreSQL mit Amazon SageMaker AI und pgvector
- Nutzen Sie pgvector und Amazon Aurora PostgreSQL für die Verarbeitung natürlicher Sprache, Chatbots und Stimmungsanalyse

Im Folgenden sind die Vorteile der Verwendung von pgvector und Aurora PostgreSQL-Compatible aufgeführt:

- Es unterstützt die exakte und ungefähre Suche nach dem nächsten Nachbarn. Es unterstützt auch die folgenden Ähnlichkeitsmetriken: L2-Entfernung, inneres Produkt und Kosinusdistanz.
- Es unterstützt die Indexierung Inverted File with Flat Compression (IVFFlat) und Hierarchical Navigable Small Worlds (HNSW).
- Sie können die Vektorsuche mit Abfragen über domänenspezifische Daten kombinieren, die in derselben PostgreSQL-Instanz verfügbar sind.
- Aurora PostgreSQL-Compatible ist für mehrstufiges Caching optimiert I/O und bietet dieses. Bei Workloads, die den verfügbaren Instanzspeicher überschreiten, kann pgvector die Abfragen pro Sekunde für die Vektorsuche um das bis zu 8-fache erhöhen.

Amazon Neptune Analytics

Amazon Neptune Analytics ist eine speicheroptimierte Graphdatenbank-Engine für Analysen. Sie unterstützt eine Bibliothek mit optimierten Algorithmen für die Graphanalyse, Grafikabfragen mit niedriger Latenz und Vektorsuchfunktionen innerhalb von Graphendurchläufen. Es verfügt auch über eine integrierte Vektorähnlichkeitssuche. Es bietet einen Endpunkt, um ein Diagramm zu erstellen, Daten zu laden, Abfragen aufzurufen und eine Vektorähnlichkeitssuche durchzuführen. Weitere Informationen zum Erstellen eines RAG-based Systems, das Neptune Analytics verwendet, finden Sie unter Verwenden von Wissensgraphen zum Erstellen von GraphRag-Anwendungen mit Amazon Bedrock und Amazon Neptune (Blogbeitrag).AWS

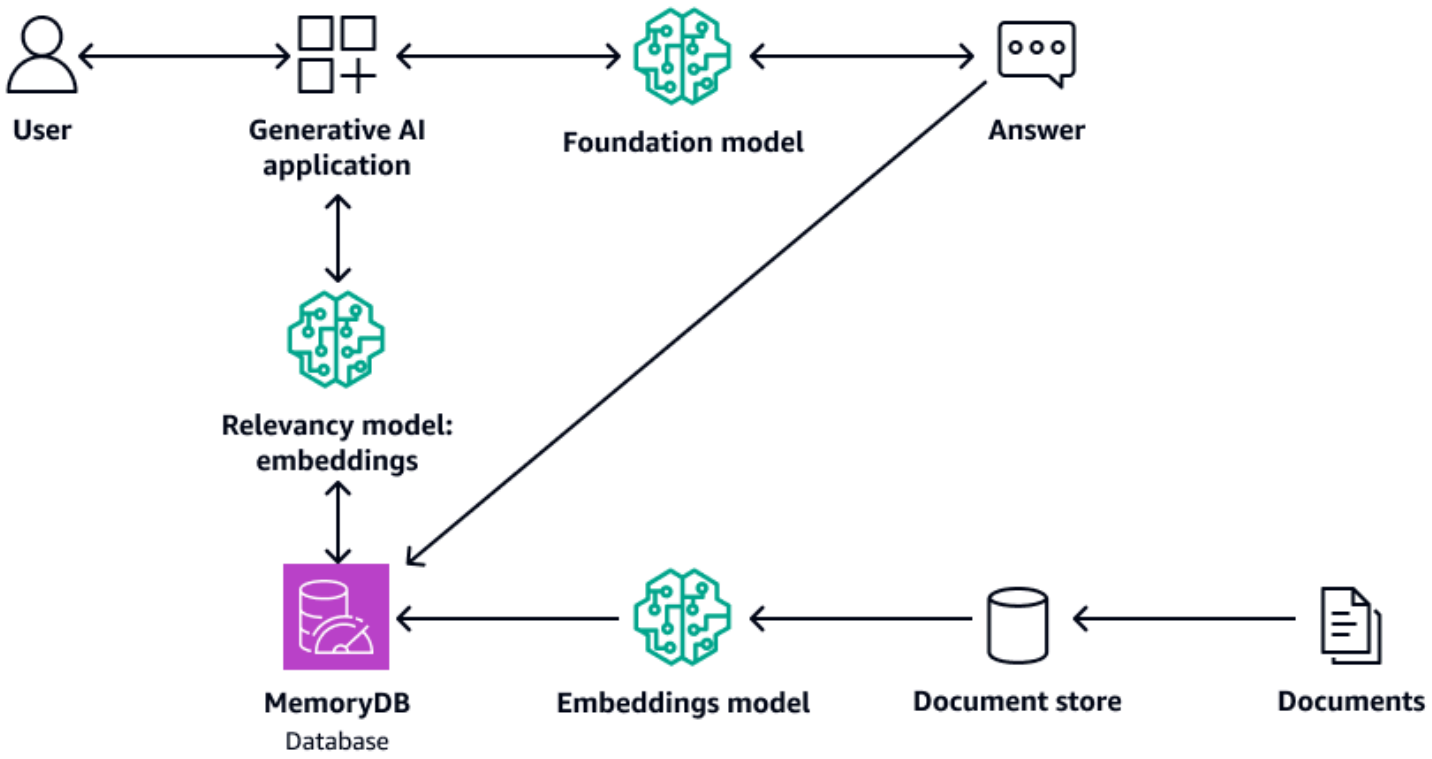
Im Folgenden sind die Vorteile der Verwendung von Neptune Analytics aufgeführt:

- Sie können Einbettungen in Grafikabfragen speichern und durchsuchen.
- Wenn Sie Neptune Analytics mit integrierenLangChain, unterstützt diese Architektur Graphabfragen in natürlicher Sprache.
- Diese Architektur speichert große Graphdatensätze im Speicher.

Amazon MemoryDB

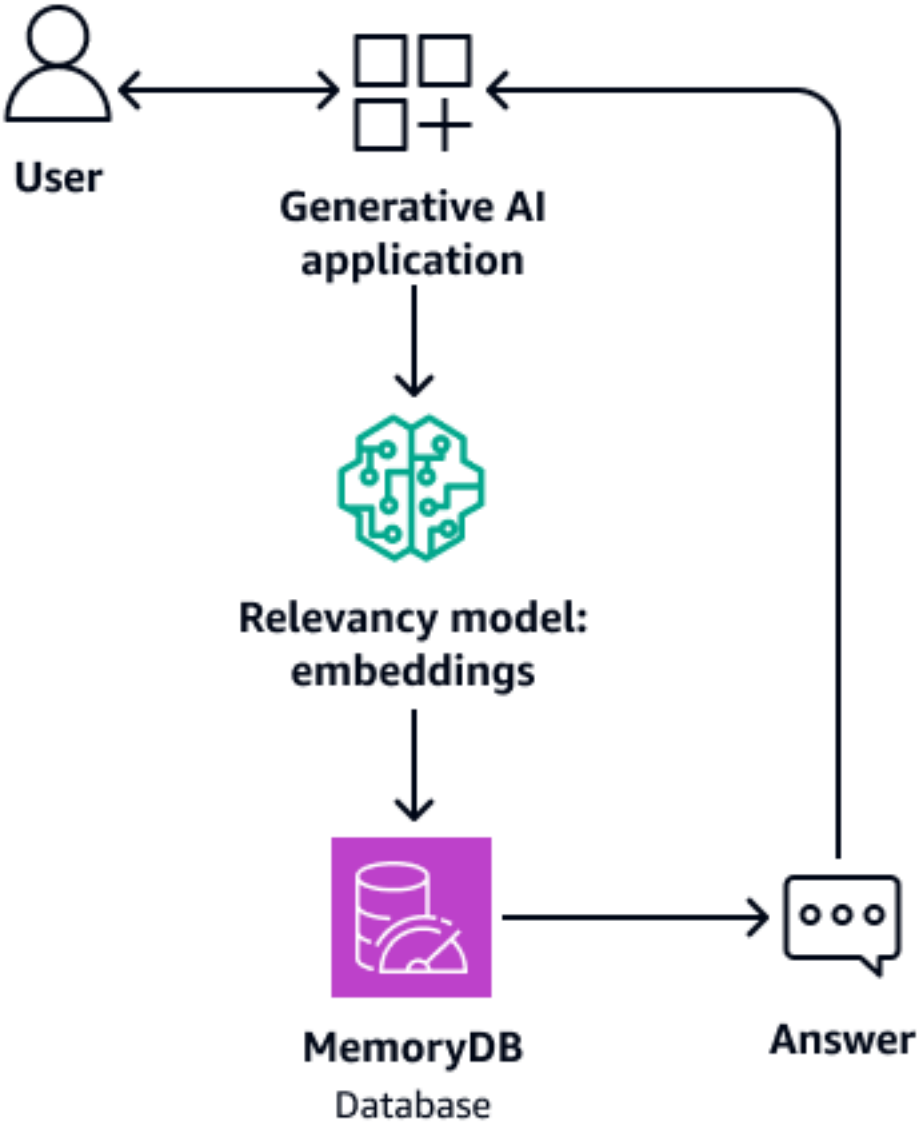
Amazon MemoryDB ist ein langlebiger In-Memory-Datenbankservice, der ultraschnelle Leistung bietet. Alle Ihre Daten werden im Speicher gespeichert, der Lesevorgänge im Mikrosekundenbereich, Schreiblatenz im einstelligen Millisekundenbereich und hohen Durchsatz unterstützt. Die Vektorsuche für MemoryDB erweitert die Funktionalität von MemoryDB und kann in Verbindung mit vorhandenen MemoryDB-Funktionen verwendet werden. Weitere Informationen finden Sie unter Fragen beantworten mit LLM und RAG-Repository auf. GitHub

Das folgende Diagramm zeigt eine Beispielarchitektur, die MemoryDB als Vektordatenbank verwendet.



Im Folgenden sind die Vorteile der Verwendung von MemoryDB aufgeführt:

- Es unterstützt sowohl Flat- als auch HNSW-Indizierungsalgorithmen. Weitere Informationen finden Sie unter Die Vektorsuche für Amazon MemoryDB ist jetzt allgemein im News-Blog verfügbar AWS
- Es kann auch als Pufferspeicher für das Foundation-Modell dienen. Dies bedeutet, dass zuvor beantwortete Fragen aus dem Puffer abgerufen werden, anstatt den Abruf- und Generierungsprozess erneut zu durchlaufen. Das folgende Diagramm zeigt diesen Prozess.



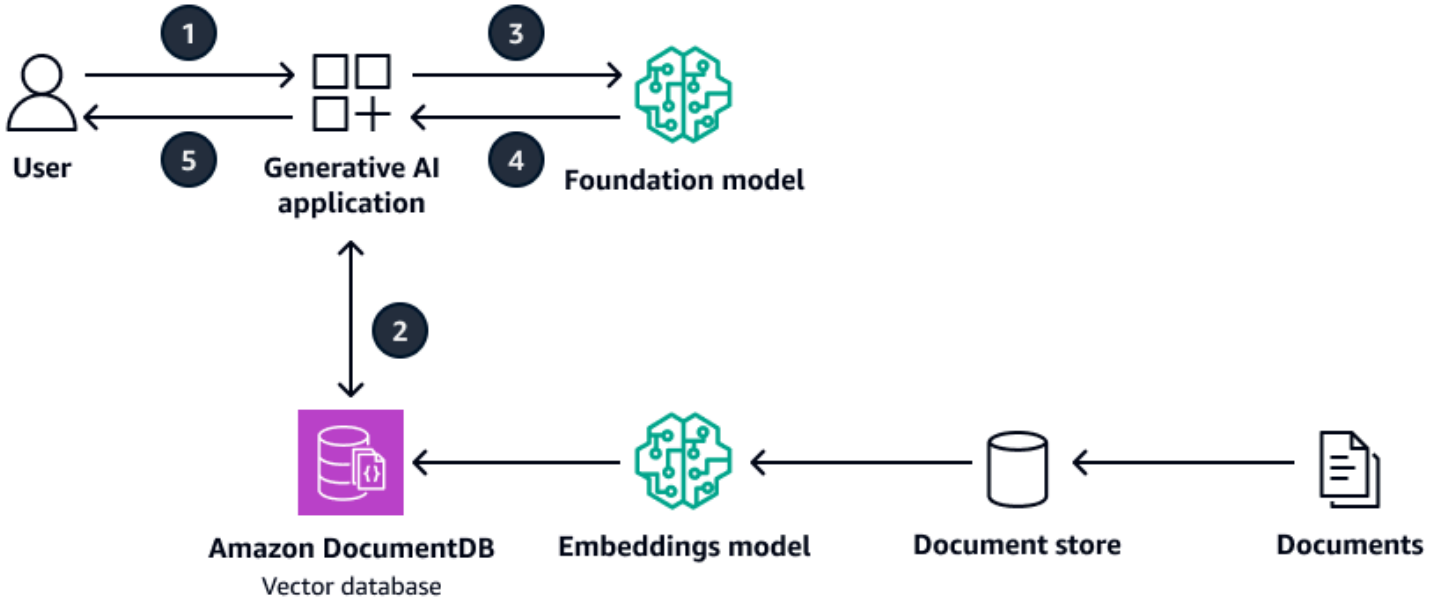
- Da sie eine In-Memory-Datenbank verwendet, bietet diese Architektur eine Abfragezeit im einstelligen Millisekundenbereich für die semantische Suche.
- Sie ermöglicht bis zu 33.000 Abfragen pro Sekunde bei einem Erinnerungsvermögen von 95—99% und 26.500 Abfragen pro Sekunde bei einem Wiedererkennungswert von mehr als 99% Weitere Informationen finden Sie im Video AWS re:Invent 2023 — Ultra-low Latenzvektorsuche für Amazon MemoryDB auf. YouTube

Amazon DocumentDB

Amazon DocumentDB (mit MongoDB-Kompatibilität) ist ein schneller, zuverlässiger und vollständig verwalteter Datenbankservice. Er macht es einfach, MongoDB kompatible Datenbanken in der Cloud

einzurichten, zu betreiben und zu skalieren. Die Vektorsuche für Amazon DocumentDB kombiniert die Flexibilität und die umfangreichen Abfragefunktionen einer JSON-based Dokumentendatenbank mit der Leistungsfähigkeit der Vektorsuche. Weitere Informationen finden Sie unter Fragen beantworten mit LLM und RAG-Repository auf. GitHub

Das folgende Diagramm zeigt eine Beispielarchitektur, die Amazon DocumentDB als Vektordatenbank verwendet.



Das Diagramm zeigt den folgenden Workflow:

1. Der Benutzer sendet eine Anfrage an die generative KI-Anwendung.
2. Die generative KI-Anwendung führt eine Ähnlichkeitssuche in der Amazon DocumentDB DocumentDB-Vektordatenbank durch und ruft die entsprechenden Dokumentauszüge ab.
3. Die generative KI-Anwendung aktualisiert die Benutzerabfrage mit dem abgerufenen Kontext und leitet die Aufforderung an das Ziel-Foundation-Modell weiter.
4. Das Foundation-Modell verwendet den Kontext, um eine Antwort auf die Frage des Benutzers zu generieren, und gibt die Antwort zurück.
5. Die generative KI-Anwendung gibt die Antwort an den Benutzer zurück.

Im Folgenden sind die Vorteile der Verwendung von Amazon DocumentDB aufgeführt:

- Es unterstützt sowohl HNSW- als auch IVFFlat-Indexierungsmethoden.

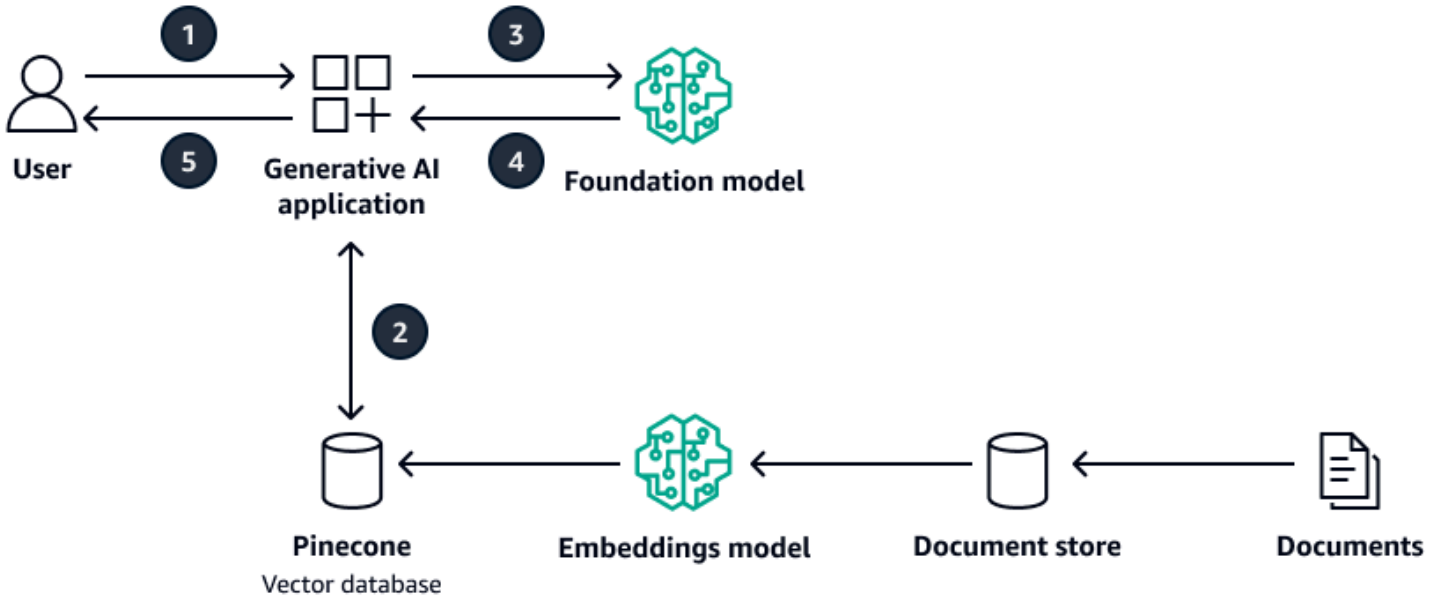
- Es unterstützt bis zu 2.000 Dimensionen in den Vektordaten und unterstützt die Abstandsmetriken Euklid, Kosinus und Punktprodukt.
- Es bietet Reaktionszeiten im Millisekundenbereich.

Tannenzapfen

Pineconeist eine vollständig verwaltete Vektordatenbank, mit der Sie Produktionsanwendungen um Vektorsuche erweitern können. Sie ist über die verfügbar AWS Marketplace. Die Abrechnung basiert auf der Nutzung. Die Gebühren werden berechnet, indem der Pod-Preis mit der Pod-Anzahl multipliziert wird. Weitere Informationen zum Aufbau eines RAG-based Systems, das Folgendes nutztPinecone, finden Sie in den folgenden AWS Blogbeiträgen:

- Reduzieren Sie Halluzinationen mithilfe von RAG mithilfe von Pinecone Vektordatenbanken und Amazon AI Llama-2 SageMaker JumpStart
- Verwenden Sie Amazon SageMaker AI Studio, um mit Llama 2 eine RAG-Lösung zur Beantwortung von Fragen zu erstellen und schnell Pinecone zu experimentieren LangChain

Das folgende Diagramm zeigt eine Beispielarchitektur, die Pinecone als Vektordatenbank verwendet wird.



Das Diagramm zeigt den folgenden Workflow:

1. Der Benutzer sendet eine Anfrage an die generative KI-Anwendung.
2. Die generative KI-Anwendung führt eine Ähnlichkeitssuche in der Pinecone Vektordatenbank durch und ruft die entsprechenden Dokumentenauszüge ab.
3. Die generative KI-Anwendung aktualisiert die Benutzerabfrage mit dem abgerufenen Kontext und sendet die Aufforderung an das Ziel-Foundation-Modell.
4. Das Foundation-Modell verwendet den Kontext, um eine Antwort auf die Frage des Benutzers zu generieren, und gibt die Antwort zurück.
5. Die generative KI-Anwendung gibt die Antwort an den Benutzer zurück.

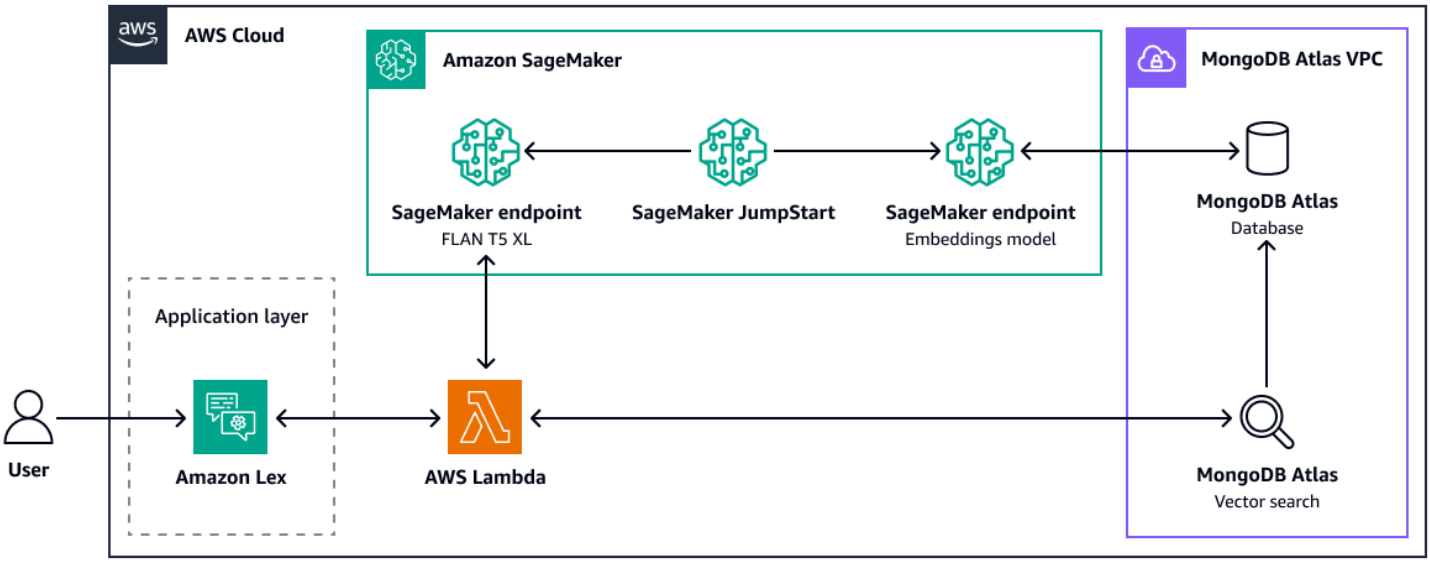
Im Folgenden sind die Vorteile der Verwendung von aufgeführtPinecone:

- Es handelt sich um eine vollständig verwaltete Vektordatenbank, mit der Sie Ihre eigene Infrastruktur nicht verwalten müssen.
- Sie bietet zusätzliche Funktionen wie Filterung, Live-Indexaktualisierungen und Keyword-Boosting (Hybridsuche).

MongoDB-Atlas

MongoDB Atlasist eine vollständig verwaltete Cloud-Datenbank, die die gesamte Komplexität der Bereitstellung und Verwaltung Ihrer Bereitstellungen bewältigt. AWS Sie können Vector Search for verwendenMongoDB Atlas, um Vektor-Einbettungen in Ihrer Datenbank zu speichern. MongoDB Amazon Bedrock Knowledge Bases unterstützt MongoDB Atlas Vektorspeicher. Weitere Informationen finden Sie in der MongoDB Dokumentation unter Erste Schritte mit der Amazon Bedrock Knowledge Base-Integration.

Weitere Informationen zur Verwendung der MongoDB Atlas Vektorsuche für RAG finden Sie unter Retrieval-Augmented Generierung mitLangChain, Amazon SageMaker AI JumpStart und MongoDB Atlas Semantic Search (AWS Blogbeitrag). Das folgende Diagramm zeigt die Lösungsarchitektur, die in diesem Blogbeitrag detailliert beschrieben wird.



Im Folgenden sind die Vorteile der Verwendung der MongoDB Atlas Vektorsuche aufgeführt:

- Sie können Ihre bestehende Implementierung von verwendenMongoDB Atlas, um Vektoreinbettungen zu speichern und zu durchsuchen.
- Sie können die MongoDBAbfrage-API verwenden, um die Vektoreinbettungen abzufragen.
- Sie können die Vektorsuche und die Datenbank unabhängig voneinander skalieren.
- Vektoreinbettungen werden in der Nähe der Quelldaten (Dokumente) gespeichert, was die Indizierungsleistung verbessert.

Weben

Weaviateist eine beliebte Open-Source-Vektordatenbank mit niedriger Latenz, die multimodale Medientypen wie Text und Bilder unterstützt. In der Datenbank werden sowohl Objekte als auch Vektoren gespeichert, wodurch die Vektorsuche mit strukturierter Filterung kombiniert wird. Weitere Informationen zur Verwendung von Weaviate Amazon Bedrock zur Erstellung eines RAG-Workflows finden Sie unter Erstellen unternehmensfähiger generativer KI-Lösungen mit Cohere Foundation-Modellen in Amazon Bedrock und Weaviate Vector Database auf AWS Marketplace (Blogbeitrag).AWS

Im Folgenden sind die Vorteile der Verwendung von: Weaviate

- Es ist Open Source und wird von einer starken Community unterstützt.
- Es ist für die Hybridsuche (sowohl Vektoren als auch Schlüsselwörter) konzipiert.

- Sie können es AWS als verwaltetes Software-as-a-Service (SaaS) -Angebot oder als Kubernetes-Cluster bereitstellen.

Generatoren für RAG-Workflows

Große Sprachmodelle (LLMs) sind sehr große Deep-Learning-Modelle, die für riesige Datenmengen vorab trainiert wurden. Sie sind unglaublich flexibel. LLMs können verschiedene Aufgaben ausführen, z. B. Fragen beantworten, Dokumente zusammenfassen, Sprachen übersetzen und Sätze vervollständigen. Sie haben das Potenzial, die Erstellung von Inhalten und die Art und Weise, wie Menschen Suchmaschinen und virtuelle Assistenten nutzen, zu stören. LLMs sind zwar nicht perfekt, weisen aber eine bemerkenswerte Fähigkeit auf, Vorhersagen auf der Grundlage einer relativ geringen Aufforderung oder Anzahl von Eingaben zu treffen.

LLMs sind eine wichtige Komponente einer RAG-Lösung. Für benutzerdefinierte RAG-Architekturen gibt es zwei, AWS-Services die als Hauptoptionen dienen:

- Amazon Bedrock ist ein vollständig verwalteter Service, der Ihnen LLMs von führenden KI-Unternehmen und Amazon über eine einheitliche API zur Verfügung stellt.
- Amazon SageMaker AI JumpStart ist ein ML-Hub, der Basismodelle, integrierte Algorithmen und vorgefertigte ML-Lösungen bietet. Mit SageMaker KI JumpStart können Sie auf vortrainierte Modelle zugreifen, einschließlich Basismodelle. Sie können auch Ihre eigenen Daten verwenden, um die vortrainierten Modelle zu optimieren.

Amazon Bedrock

Amazon Bedrock bietet branchenführende Modelle vonAnthropic,,Stability AI,Meta, Cohere AI21 LabsMistral AI, und Amazon. Eine vollständige Liste finden Sie unter Unterstützte Foundation-Modelle in Amazon Bedrock. Mit Amazon Bedrock können Sie Modelle auch mit Ihren eigenen Daten anpassen.

Sie können die Modellleistung bewerten, um festzustellen, welche Modelle für Ihren RAG-Anwendungsfall am besten geeignet sind. Sie können die neuesten Modelle testen und auch testen, welche Funktionen und Funktionen die besten Ergebnisse liefern und das zum besten Preis. Das Anthropic Claude Sonnet-Modell wird häufig für RAG-Anwendungen verwendet, da es sich bei einer Vielzahl von Aufgaben hervorragend eignet und ein hohes Maß an Zuverlässigkeit und Vorhersagbarkeit bietet.

SageMaker AI JumpStart

SageMaker KI JumpStart bietet vortrainierte Open-Source-Modelle für eine Vielzahl von Problemtypen. Sie können diese Modelle vor der Bereitstellung schrittweise trainieren und optimieren. Sie können auf die vortrainierten Modelle, Lösungsvorlagen und Beispiele über die SageMaker JumpStart KI-Landingpage in Amazon SageMaker AI Studio zugreifen oder das SageMaker AI Python SDK verwenden.

SageMaker KI JumpStart bietet hochmoderne Basismodelle für Anwendungsfälle wie das Schreiben von Inhalten, Codegenerierung, Beantwortung von Fragen, Verfassen von Texten, Zusammenfassung, Klassifizierung, Informationsabruf und mehr. Verwenden Sie JumpStart Basismodelle, um Ihre eigenen generativen KI-Lösungen zu erstellen und maßgeschneiderte Lösungen mit zusätzlichen SageMaker KI-Funktionen zu integrieren. Weitere Informationen finden Sie unter Erste Schritte mit Amazon SageMaker AI JumpStart.

SageMaker KI integriert JumpStart und verwaltet öffentlich verfügbare Basismodelle, auf die Sie zugreifen, sie anpassen und in Ihre ML-Lebenszyklen integrieren können. Weitere Informationen finden Sie unter Öffentlich verfügbare Basismodelle. SageMaker KI umfasst JumpStart auch proprietäre Basismodelle von Drittanbietern. Weitere Informationen finden Sie unter Proprietäre Basismodelle.

Wählen Sie eine Option zum Abrufen erweiterter Generierung auf AWS

In den Abschnitten Vollständig verwaltete RAG-Optionen und Benutzerdefinierte RAG-Architekturen dieses Handbuchs werden verschiedene Ansätze zum Aufbau einer RAG-basierten Suchlösung beschrieben. AWS In diesem Abschnitt wird beschrieben, wie Sie je nach Anwendungsfall zwischen diesen Optionen wählen können. In einigen Situationen funktioniert möglicherweise mehr als eine Option. In diesem Szenario hängt die Wahl von der Einfachheit der Implementierung, den in Ihrer Organisation verfügbaren Fähigkeiten und den Richtlinien und Standards Ihres Unternehmens ab.

Wir empfehlen Ihnen, die vollständig verwalteten und benutzerdefinierten RAG-Optionen in der folgenden Reihenfolge zu betrachten und die erste Option auszuwählen, die zu Ihrem Anwendungsfall passt:

1. Verwenden Sie Amazon Q Business, es sei denn:
   - Dieser Service ist in Ihrem Land nicht verfügbar AWS-Region, und Ihre Daten können nicht in eine Region verschoben werden, in der er verfügbar ist
   - Sie haben einen bestimmten Grund, den RAG-Workflow anzupassen
   - Sie möchten eine bestehende Vektordatenbank oder ein bestimmtes LLM verwenden
2. Verwenden Sie Wissensdatenbanken für Amazon Bedrock, es sei denn:
   - Sie haben eine Vektordatenbank, die nicht unterstützt wird
   - Sie haben einen bestimmten Grund, den RAG-Workflow anzupassen
3. Kombinieren Sie Amazon Kendra mit einem Generator Ihrer Wahl, es sei denn:
   - Sie möchten Ihre eigene Vektordatenbank wählen
   - Sie möchten die Chunking-Strategie anpassen
4. Wenn Sie mehr Kontrolle über den Retriever haben und Ihre eigene Vektordatenbank auswählen möchten:
   - Wenn Sie nicht über eine bestehende Vektordatenbank verfügen und keine Abfragen mit niedriger Latenz oder Diagrammen benötigen, sollten Sie Amazon OpenSearch Service in Betracht ziehen.
   - Wenn Sie über eine bestehende PostgreSQL Vektordatenbank verfügen, sollten Sie die Amazon Aurora PostgreSQL and-Option in Betracht ziehen. pgvector

- Wenn Sie eine niedrige Latenz benötigen, sollten Sie eine In-Memory-Option wie Amazon MemoryDB oder Amazon DocumentDB in Betracht ziehen.
- Wenn Sie die Vektorsuche mit einer Grafikabfrage kombinieren möchten, sollten Sie Amazon Neptune Analytics in Betracht ziehen.
- Wenn Sie bereits eine Vektordatenbank eines Drittanbieters verwenden oder einen bestimmten Vorteil aus einer solchen Datenbank ziehen, sollten Sie PineconeMongoDB Atlas, und in Betracht ziehen. Weaviate

5. Wenn Sie sich für ein LLM entscheiden möchten:
   - Wenn Sie Amazon Q Business verwenden, können Sie den LLM nicht wählen.
   - Wenn Sie Amazon Bedrock verwenden, können Sie eines der unterstützten Foundation-Modelle wählen.
   - Wenn Sie Amazon Kendra oder eine benutzerdefinierte Vektordatenbank verwenden, können Sie einen der in diesem Handbuch beschriebenen Generatoren oder ein benutzerdefiniertes LLM verwenden.

   > **Note**
   >
   > Sie können Ihre benutzerdefinierten Dokumente auch verwenden, um ein vorhandenes LLM zu optimieren, um die Genauigkeit der Antworten zu erhöhen. Weitere Informationen finden Sie unter Vergleich von RAG und Feinabstimmung in diesem Handbuch.

6. Wenn Sie bereits über eine Implementierung von Amazon SageMaker AI Canvas verfügen, die Sie verwenden möchten, oder wenn Sie RAG-Antworten verschiedener Anbieter vergleichen möchten LLMs, sollten Sie Amazon SageMaker AI Canvas in Betracht ziehen.

Schlussfolgerung

In diesem Handbuch werden die verschiedenen Optionen beschrieben, auf denen ein Retrieval Augmented Generation (RAG) -System aufgebaut werden kann. AWS Sie können mit vollständig verwalteten Services wie Amazon Q Business und Amazon Bedrock Knowledge Bases beginnen. Wenn Sie mehr Kontrolle über den RAG-Workflow haben möchten, können Sie einen benutzerdefinierten Retriever wählen. Für einen Generator können Sie eine API verwenden, um ein unterstütztes LLM in Amazon Bedrock aufzurufen, oder Sie können Ihr eigenes LLM mithilfe von Amazon AI bereitstellen. SageMaker JumpStart Lesen Sie die Empfehlungen unter Auswahl einer RAG-Option, um herauszufinden, welche Option für Ihren Anwendungsfall am besten geeignet ist. Nachdem Sie die beste Option für Ihren Anwendungsfall ausgewählt haben, verwenden Sie die Referenzen in diesem Handbuch, um mit der Erstellung Ihrer RAG-basierten Anwendung zu beginnen.

Dokumentverlauf

In der folgenden Tabelle werden wichtige Änderungen in diesem Leitfaden beschrieben. Um Benachrichtigungen über zukünftige Aktualisierungen zu erhalten, können Sie einen RSS-Feed abonnieren.

Änderung	Beschreibung	Datum
Erste Veröffentlichung	—	28. Oktober 2024

Die vorliegende Übersetzung wurde maschinell erstellt. Im Falle eines Konflikts oder eines Widerspruchs zwischen dieser übersetzten Fassung und der englischen Fassung (einschließlich infolge von Verzögerungen bei der Übersetzung) ist die englische Fassung maßgeblich.